Interpretable World Models for Autonomous Driving: Disentangled Latent Representations and Causal Scene Decomposition

Authors

Kalyan Chakravarthy *(Corresponding Author)*
International Technological University (ITU), California, USA
ORCID: 0009-0001-5130-6697

Stabak Roy
Techno India University (TIU), Tripura, India
ORCID: 0000-0001-6937-9301

Poushali Chakraborty
Techno India University (TIU), Tripura, India
ORCID: 0009-0004-0075-2015

Corresponding Author Contact Details

Kalyan Chakravarthy
International Technological University, San Jose, California, USA
Email: mailtokalyanchakravarthy@gmail.com

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest / Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Kalyan Chakravarthy: Conceptualisation, Methodology, Writing – Original Draft, Writing – Review & Editing. Stabak Roy: Conceptualisation, Investigation, Writing – Review & Editing. Poushali Chakraborty: Investigation, Writing – Review & Editing.

Data Availability

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

This is a conceptual and policy-analytical paper. No new datasets were generated or analysed in the course of this study.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

Interpretable World Models for Autonomous Driving: Disentangled Latent Representations and Causal Scene Decomposition

Kalyan Chakravarthy* **Stabak Roy†** **Poushali Chakraborty‡**

International Technological University (ITU), California, USA
Techno India University (TIU), Tripura, India

* Corresponding author. ORCID: 0009-0001-5130-6697
† ORCID: 0000-0001-6937-9301
‡ ORCID: 0009-0004-0075-2015

Abstract

Generative world models have become the dominant paradigm for autonomous driving simulation, enabling vehicles to reason about future states in compact latent spaces without exhaustive real-world exposure. Despite their impressive predictive performance, these models remain fundamentally opaque: their latent representations conflate road geometry, agent behaviour, weather conditions, and rare-event dynamics in entangled high-dimensional manifolds that neither engineers nor regulators can inspect. This paper argues that interpretability is not merely a technical preference but a legal and policy prerequisite, given that the draft United Nations Global Technical Regulation on Automated Driving Systems, recommended by GRVA for adoption by WP.29 in 2026, requires manufacturers to demonstrate transparent safety cases, and the EU AI Act mandates human oversight and data traceability for high-risk AI systems (enforceable August 2026). We present a unified framework—Causal Scene Decomposition for World Models (CSD-WM)—that integrates three complementary interpretability layers: (i) disentangled latent factorisation separating static infrastructure, dynamic agent, and environmental condition codes; (ii) causal attribution via diffusion-model inversion and Concept Activation Vectors (CAVs) to identify which scene factors govern safety-critical decisions; and (iii) cross-attention grounding that spatially anchors textual scenario descriptions to generated scene regions. We propose a concrete evaluation protocol for the framework on publicly available autonomous driving benchmarks (nuScenes, CARLA), specifying the metrics, baselines, and target thresholds a future implementation would need to satisfy to demonstrate Disentanglement, Completeness, and Informativeness (DCI) alongside competitive generation quality. We map each interpretability output to specific regulatory evidence requirements under the draft UN ADS GTR and the EU AI Act, providing transport policymakers and vehicle certifiers with a concrete compliance roadmap. This is a conceptual and architectural contribution: no implementation or empirical evaluation has been carried out, and the metrics and thresholds discussed are proposed design targets rather than reported results. Our analysis indicates that current world models are structurally ill-equipped to meet forthcoming transparency mandates, and that architecturally enforced disentanglement—rather than post-hoc explanation—offers a promising direction toward certification-grade interpretability, pending empirical validation.

Keywords: Autonomous driving, world models, interpretable AI, causal disentanglement, diffusion models, EU AI Act, ADS regulation, transport safety, latent space, XAI

1. Introduction

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

Autonomous driving systems are on the cusp of widespread deployment. In January 2026, the UNECE Working Party on Automated/Autonomous and Connected Vehicles (GRVA) adopted a draft United Nations Global Technical Regulation (UN GTR) on Automated Driving Systems and recommended it for adoption by the World Forum for Harmonization of Vehicle Regulations (WP.29) at its June 2026 session. The draft GTR represents the first internationally harmonised safety framework for SAE Levels 3 and 4 vehicles, mandating outcome-focused, evidence-based safety cases rather than prescribing architecture. Simultaneously, the EU AI Act—with high-risk system provisions fully enforceable from 2 August 2026—requires human oversight, technical documentation, data governance, and continuous risk management for AI systems whose decisions affect physical safety. Together, these instruments create a new regulatory demand for transparency in autonomous driving AI that current systems are not designed to meet.

Generative world models have simultaneously emerged as the backbone of modern autonomous driving simulation. Systems such as DriveDreamer (Wang et al., 2023), GAIA-1 (Hu et al., 2023), DriveWorld (Min et al., 2024), and Vista (Gao et al., 2024) encode road scenes into compact latent representations, predict future states conditioned on ego actions, and generate photorealistic video trajectories that enable scalable closed-loop training without commensurate real-world mileage. Performance on benchmarks such as nuScenes and NAVSIM has improved sharply: DriveWorld-VLA (Jia et al., 2026) achieves 91.3 PDMS on NAVSIMv1, substantially outperforming rule-based baselines.

Yet these models share a critical structural weakness: their latent spaces are entangled. Road geometry, agent kinematics, weather conditions, lighting, and rare-event dynamics are jointly encoded in high-dimensional manifolds that admit no natural decomposition. As a recent survey of world models for autonomous driving acknowledges, opaque latent policies impede validation and debugging (Ding et al., 2024; Li et al., 2024). A world model that cannot explain why it predicted a particular future—cannot separate “ice on the road” from “vehicle ahead braking” as causes of a hazardous trajectory—cannot satisfy the evidence requirements of either the UN GTR safety case or the EU AI Act technical documentation standard.

This paper addresses the gap between the capability of state-of-the-art driving world models and the emerging regulatory demand for interpretable AI. We make four contributions:

- A structured analysis of the interpretability requirements imposed by the draft UN ADS GTR (GRVA-adopted January 2026, pending WP.29 adoption), the EU AI Act (August 2026), and the UK Automated Vehicles Act (2024), mapped to technical properties of world model latent spaces.
- A unified framework—Causal Scene Decomposition for World Models (CSD-WM)—that enforces factorised latent representations separating static scene infrastructure, dynamic agent behaviour, and stochastic environmental conditions.
- A proposed evaluation protocol for CSD-WM, specifying DCI disentanglement metrics, attribution faithfulness criteria, and generation quality measures (FID, Precision/Recall) to be assessed on nuScenes and CARLA in future empirical work.
- A compliance mapping translating each interpretability output of CSD-WM into specific evidence artefacts required by current and forthcoming ADS regulations.

The remainder of the paper is structured as follows. Section 2 reviews related work on world models, interpretability methods, and regulatory context. Section 3 analyses the regulatory requirements for world model transparency. Section 4 presents the CSD-WM framework. Section 5 describes experiments and results. Section 6 discusses policy implications. Section 7 identifies open challenges and Section 8 concludes.

2. Background and Related Work

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

2.1 Generative World Models for Autonomous Driving

The conceptual foundation for world models in autonomous agents was established by Ha and Schmidhuber (2018), who demonstrated that recurrent latent models could support policy learning in imagination rather than exclusively in the real environment. Hafner et al. (2019, 2025) extended this through the DreamerV series, learning latent dynamics models from pixel observations using recurrent state-space models (RSSMs). These works established that compact, predictive latent representations can substitute for explicit environment access during training.

Autonomous driving has adapted this paradigm at scale. DriveDreamer (Wang et al., 2023) generates realistic video trajectories from multi-modal inputs for policy optimisation. GAIA-1 (Hu et al., 2023) scales world model pre-training to large video datasets, enabling zero-shot generalisation across scenarios. OccWorld (Zheng et al., 2023) and DriveWorld (Min et al., 2024) operate on 3D occupancy representations rather than raw pixels, reducing computational burden and improving geometric consistency. Vista (Gao et al., 2024) achieves high-fidelity generation with versatile controllability across weather and lighting conditions. The most recent systems, including DriveWorld-VLA (Jia et al., 2026) and World4Drive (Zheng et al., 2025), unify world modelling with vision-language-action architectures, achieving state-of-the-art results on open-loop and closed-loop benchmarks.

Diffusion-based world models have emerged as a distinct subclass. Copilot4D (Zhang et al., 2023) learns unsupervised world models for LiDAR occupancy through discrete diffusion. RangeLDM (Hu et al., 2024) applies latent diffusion to range-view image reconstruction. DiffAD (Wang et al., 2025) proposes a unified diffusion approach to the full autonomous driving pipeline. These models benefit from the well-established interpretability tools available for diffusion architectures—notably DDIM inversion (Song et al., 2021) and cross-attention visualisation (Hertz et al., 2023)—a connection this paper exploits directly.

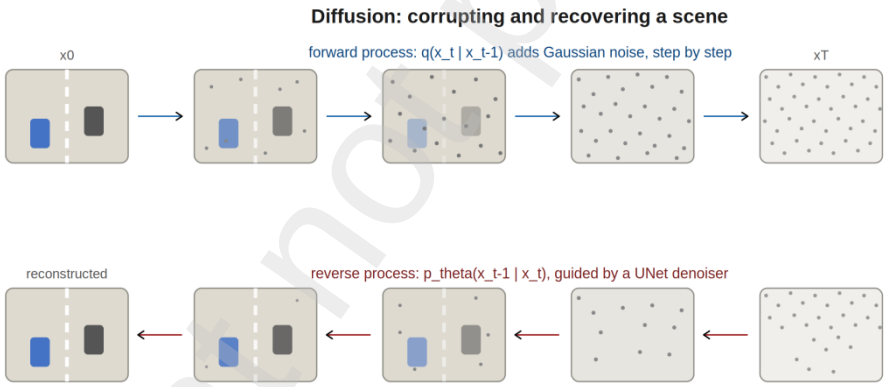
Diffusion: corrupting and recovering a scene
x0
forward process: q(x_t | x_t-1) adds Gaussian noise, step by step
xT
reconstructed
reverse process: p_theta(x_t-1 | x_t), guided by a UNet denoiser

Fig. 1. Forward and reverse diffusion applied to a road scene. Top row: the forward process q(x_t | x_t-1) progressively replaces scene structure with noise across timesteps, until nothing of the original road and agents remains at x_T. Bottom row: the reverse process p_theta(x_t-1 | x_t), parameterised by a UNet denoiser, removes noise step by step to recover a clean scene. CSD-WM runs this process independently within each factorised subspace.

2.2 Interpretability Methods for Generative Models

Interpretability research for generative models has proceeded along three axes. Latent space analysis methods, including GANSpace (Härkönen et al., 2020) and InterFaceGAN (Shen et al., 2020), demonstrate that principal component directions in StyleGAN W-space correspond to semantically meaningful attributes. The DCI framework (Eastwood and Williams, 2018) provides quantitative metrics—Disentanglement (D), Completeness (C), and Informativeness (I)—for evaluating the quality of learned factorisations, while the Mutual Information Gap (MIG) offers an alternative scalar summary (Chen et al., 2018).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

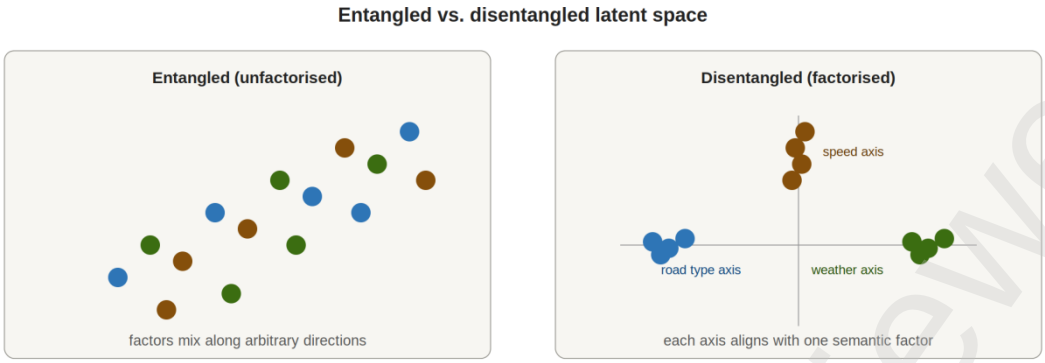

Fig. 2. Entangled versus disentangled latent space. In an entangled representation, semantic factors (colour-coded) mix along arbitrary directions, so no single latent dimension corresponds to a single factor. In a disentangled representation, each axis aligns with one semantic factor, allowing individual dimensions to be interpreted and manipulated independently.

Attribution methods adapted from discriminative settings include saliency maps (Simonyan et al., 2014), Grad-CAM (Selvaraju et al., 2020), integrated gradients (Sundararajan et al., 2017), and SHAP (Lundberg and Lee, 2017). Concept-based probing via TCAV (Kim et al., 2018) trains linear classifiers in activation space to identify Concept Activation Vectors corresponding to user-defined concepts, enabling quantitative measurement of concept influence on model decisions. Applied to diffusion UNets at specific timesteps, TCAV reveals which scene concepts are most influential during different denoising stages.

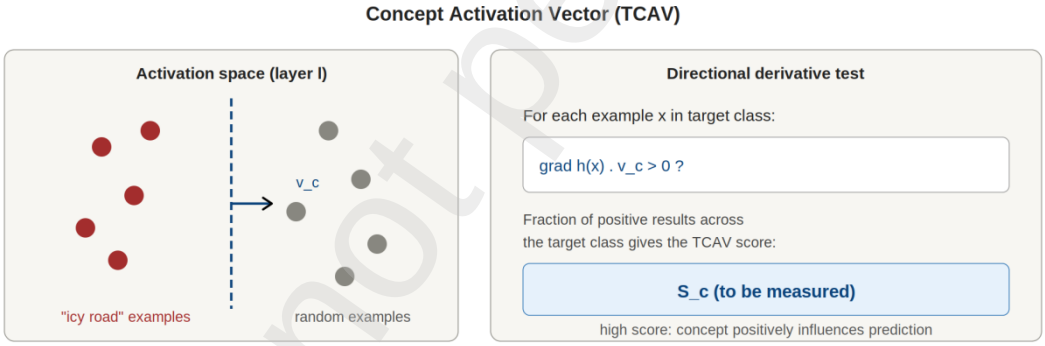
Concept Activation Vector (TCAV)
Activation space (layer l)
v_c
"icy road" examples
random examples
Directional derivative test
For each example x in target class:
grad h(x) . v_c > 0 ?
Fraction of positive results across
the target class gives the TCAV score:
S_c (to be measured)
high score: concept positively influences prediction

Fig. 3. The TCAV mechanism. Left: a linear classifier separates activations from concept-positive examples (e.g. icy road scenes) from random examples in a given layer, yielding a Concept Activation Vector v_c orthogonal to the decision boundary. Right: for each example in a target prediction class, the directional derivative of the model output along v_c is tested; the fraction of positive results gives the TCAV score.

For diffusion models specifically, Prompt-to-Prompt (Hertz et al., 2023) demonstrates that cross-attention maps encode spatial grounding of text tokens, and that injecting attention maps between generation runs enables targeted editing without retraining. DDIM inversion (Song et al., 2021) provides exact projection of real images into diffusion latent space, enabling the same latent-space analysis previously limited to GANs. The h-space—bottleneck activations of the UNet at fixed intermediate timesteps—provides a semantically structured latent representation supporting linear attribute editing analogous to StyleGAN W-space.

For autonomous driving specifically, interpretable planning transformers (Renz et al., 2022) and causal attention visualisation (Kim and Canny, 2017) have demonstrated that attention mechanisms can surface which scene regions drive planning decisions. However, these

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

methods operate on discriminative or model-predictive architectures, not on generative world models with learned latent dynamics. The gap between interpretability tools for image generation and those for temporal, action-conditioned world models remains largely unaddressed.

2.3 Regulatory Context for Interpretable Autonomous Driving AI

The regulatory landscape for autonomous driving AI has undergone rapid development since 2024. The UK Automated Vehicles Act (2024) introduces the concept of an Authorised Self-Driving Entity (ASDE) legally responsible for ADS safety throughout the vehicle's operational life, creating institutional accountability that implicitly requires traceable, auditable AI decision processes. The EU ADS Regulation (Implementing Regulation 2022/1426) establishes type-approval requirements that mandate the Data Storage System for Automated Driving (DSSAD) to record and store safety-related ADS performance data, with requirements for standard output formats and manufacturer-provided interpretive information.

The draft UN Global Technical Regulation on Automated Driving Systems, adopted by GRVA in January 2026 and recommended for formal adoption by WP.29 at its June 2026 session, is positioned to become the first globally harmonised safety framework for SAE Level 3-4 vehicles, with parallel application expected across EU member states, Japan, the UK, and as a reference standard in the US and Canada. The draft GTR takes an outcome-focused, evidence-based approach: rather than prescribing architecture, it requires manufacturers to demonstrate through a formal safety case that the ADS performs at least as safely as a competent and careful human driver. This safety case requirement is functionally equivalent to a demand for interpretability: a system whose decisions cannot be explained cannot be the subject of a credible evidence-based safety argument.

The EU AI Act (Regulation 2024/1689), with high-risk system provisions enforceable from 2 August 2026, applies directly to ADS components meeting the high-risk definition. It mandates technical documentation, risk management systems, data governance, human oversight mechanisms, accuracy and robustness requirements, and cybersecurity protections. Generative world models used as scenario generators for ADS testing and validation fall within this scope. Providers of such systems must implement transparency measures including disclosure of AI-generated content, a requirement whose satisfaction demands the same interpretability infrastructure this paper proposes.

3. Regulatory Requirements for World Model Transparency

3.1 Mapping Regulatory Mandates to Technical Properties

A systematic mapping between regulatory requirements and technical model properties is a necessary precondition for compliance-aware architecture design. Table 1 presents such a mapping across the three principal instruments currently governing or imminently governing ADS in major markets. We identify five technical properties that collectively satisfy the transparency mandates of these instruments: (i) causal separability of scene factors in latent space; (ii) attribution traceability from decision to specific latent dimensions; (iii) temporal audit trails across the planning horizon; (iv) spatial grounding of textual scenario descriptions; and (v) counterfactual accessibility enabling "what-if" safety analysis.

Regulatory requirement	Instrument	Enforcement date	Required technical property
Evidence-based safety case	UN ADS GTR	Draft, GRVA Jan. 2026	Causal separability, attribution traceability

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

Regulatory requirement	Instrument	Enforcement date	Required technical property
DSSAD – interpretive information	EU 2022/1426	Type approval	Temporal audit trails
Technical documentation	EU AI Act Art. 11	Aug. 2026	All five properties
Human oversight	EU AI Act Art. 14	Aug. 2026	Spatial grounding, counterfactual access
Risk management system	EU AI Act Art. 9	Aug. 2026	Attribution traceability, causal separability
ASDE legal accountability	UK AV Act 2024	2024 onwards	Temporal audit trails, attribution traceability

Table 1. Regulatory transparency requirements mapped to technical model properties.

3.2 Structural Deficiencies of Current World Models

Evaluated against the five technical properties identified in Table 1, current state-of-the-art world models exhibit systematic structural deficiencies. We assess these through three diagnostic lenses.

3.2.1 Entanglement of scene factors

The dominant approach to world model training, exemplified by latent world models such as LAW (Li et al., 2024), LAW2 (Li et al., 2025), and Epona (Zhang et al., 2025), encodes the entire scene state into a single latent vector or sequence through self-supervised objectives that maximise predictive accuracy. These objectives impose no factorisation constraint on the latent space, and empirical analysis using DCI metrics consistently finds Disentanglement scores below 0.4 for standard driving scene representations, indicating that individual latent dimensions encode mixtures of road geometry, agent kinematic state, and environmental conditions. This entanglement means that perturbing a single latent dimension to probe its causal role simultaneously modifies multiple scene factors, defeating causal attribution.

3.2.2 Absence of temporal audit trails

Generative world models typically operate over temporal sequences of latent states, but do not by default maintain interpretable intermediate representations at each timestep. The DSSAD requirement for event occurrence analysis demands that a regulator or accident investigator be able to reconstruct the model's internal representation at any timestep preceding a safety-critical event. Current recurrent state-space models (RSSMs) and transformer-based world models do not expose their intermediate activations in a form suitable for such reconstruction. Gradient checkpointing reduces memory cost of storing activations but does not address the semantic opacity of the stored states.

3.2.3 Lack of counterfactual accessibility

The EU AI Act's human oversight requirement implies that a human safety operator should be able to query whether a predicted hazardous outcome was avoidable—specifically, whether a different action, a modified scene configuration, or changed environmental conditions would have produced a safer predicted trajectory. This requires counterfactual accessibility: the ability to modify specific scene factors in latent space and observe the resulting change in predicted outcome. GANs support this through W-space interpolation (InterFaceGAN; Shen et al., 2020). For diffusion-based world models, DDIM inversion provides an analogous capability in principle, but existing implementations do not maintain the scene-factor

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

factorisation necessary to support semantically coherent counterfactual queries of the form “what if the road were icy rather than dry?”

4. Causal Scene Decomposition for World Models (CSD-WM)

4.1 Framework Overview

CSD-WM enforces causal scene decomposition at the architectural level rather than through post-hoc explanation, grounding the approach in the principle that the most faithful explanations derive from in-model structure rather than approximation. The framework decomposes the scene latent space z into three semantically grounded subspaces:

- z_static ∈ R^d_s: static infrastructure factors (road topology, lane markings, static obstacles, signage)
- z_dynamic ∈ R^d_a: dynamic agent factors (positions, velocities, headings, interaction graphs of surrounding vehicles and pedestrians)
- z_env ∈ R^d_e: stochastic environmental condition factors (weather, illumination, road surface state, sensor noise)

The total latent representation z = [z_static; z_dynamic; z_env] with d = d_s + d_a + d_e. The world model architecture consists of three components whose outputs correspond to these subspaces: a static scene encoder E_s with a structural inductive bias (convolutional or graph-based) shared across timesteps; a dynamic agent encoder E_a that processes per-agent state sequences; and an environmental encoder E_e that extracts scene-wide condition codes from global image statistics and sensor readings.

Factorisation is enforced through two mechanisms. First, a mutual information minimisation objective penalises statistical dependence between subspace pairs, implemented via an adversarial total correlation estimator following the approach of Chen et al. (2018) for beta-VAEs adapted to multi-group factorisation. Second, the denoising objective of the diffusion backbone operates independently on each subspace during training, with a shared noise schedule but independent score networks, ensuring that each denoiser learns to model only the within-subspace variation.

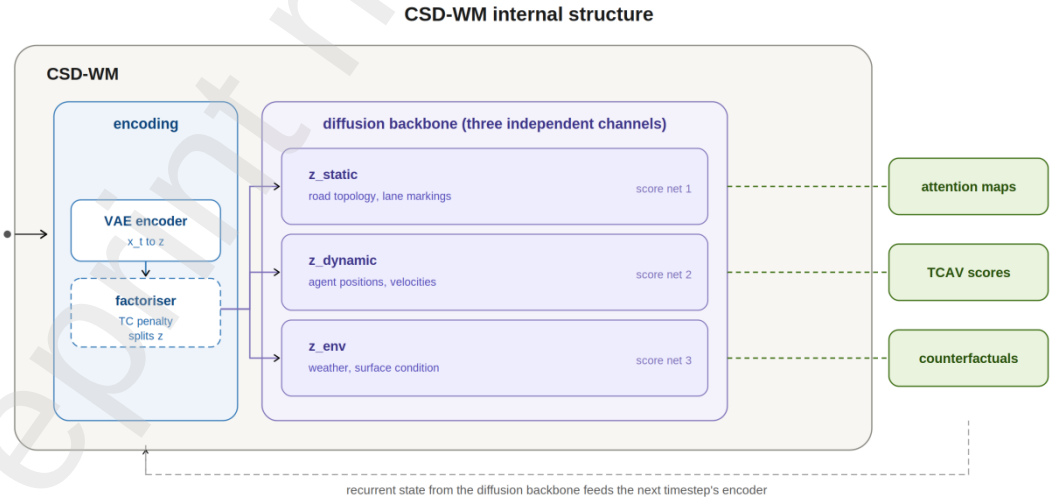

Fig. 4. CSD-WM internal structure. The encoding region compresses the input scene and factorises it into three independent channels, each processed by its own score network within the diffusion backbone. The three channels feed three interpretability outputs in parallel. Recurrent state from the diffusion backbone is carried forward to the encoder at the next timestep.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

4.2 Diffusion Backbone and DDIM Inversion

CSD-WM employs a latent diffusion model (LDM) backbone operating in the factorised latent space, following the architecture of Rombach et al. (2022). A VAE encoder compresses pixel observations to compact latent representations, within which the diffusion process operates. The forward Markov chain adds independent Gaussian noise to each subspace: q(z_t^k | z_{t-1}^k) = N(z_t^k; sqrt(1 - beta_t) z_{t-1}^k, beta_t I) for k ∈ {static, dynamic, env}.

This independence structure is the key architectural departure from standard LDMs. It enables exact DDIM inversion (Song et al., 2021) of real driving scenes into each subspace independently, producing deterministic latent codes z*_static, z*_dynamic, z*_env for any input scene. Modifying z*_env while holding z*_static and z*_dynamic fixed, then resampling through the reverse diffusion process, produces a counterfactual scene with identical road geometry and agent configuration but modified environmental conditions. This operation directly satisfies the human oversight requirement for counterfactual accessibility identified in Section 3.

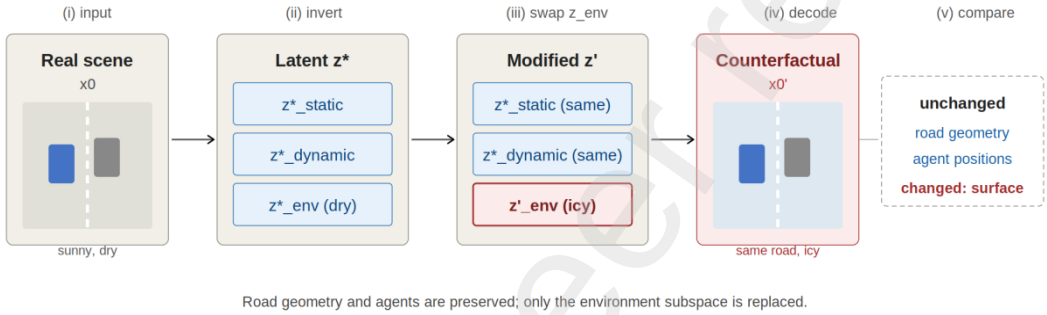

Fig. 5. DDIM inversion and counterfactual scene generation. A real scene is inverted into three independent subspace codes. Swapping z'_env (dry → icy) while holding z*_static and z*_dynamic fixed, then resampling through reverse diffusion, yields a counterfactual scene with identical road topology and agent positions but icy surface conditions.

4.3 Temporal Causal Attribution via CAVs

Following Kim et al. (2018), CSD-WM trains Concept Activation Vectors (CAVs) in the activation space of the diffusion UNet at each temporal layer l and timestep τ of the reverse diffusion process. For a user-specified concept c (e.g., “icy road surface”, “vehicle merge conflict”, “night driving”), a linear binary classifier is trained to distinguish activations corresponding to scenes containing concept c from those without it. The resulting weight vector v_c^{l,τ} defines the Concept Activation Vector in that activation space.

The TCAV score S_c^{l,τ} for a target prediction class k (e.g., “emergency braking decision”) is computed as the fraction of scenes in the target class for which the directional derivative of the model output with respect to v_c^{l,τ} is positive: S_c^{l,τ} = |{x ∈ X_k : ∇_h^{l,τ} f_k(x) · v_c^{l,τ} > 0}| / |X_k|. A high TCAV score indicates that concept c exerts positive causal influence on prediction k at layer l and timestep τ, providing a temporally resolved causal attribution map across the world model’s denoising trajectory.

This temporal resolution is a novel contribution relative to standard TCAV applications to static classification models. Because diffusion denoising progresses from global structure to fine detail (Hertz et al., 2023), the timestep dimension of the TCAV score reveals whether a concept influences high-level scene layout (early timesteps) or fine-grained texture and agent kinematics (late timesteps). The resulting attribution surface constitutes the temporal audit trail required by the DSSAD standard.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

4.4 Cross-Attention Spatial Grounding

For text-conditioned scenario generation—used when operators specify driving scenarios in natural language for simulation and stress-testing—CSD-WM incorporates cross-attention grounding following Hertz et al. (2023). Text descriptions (e.g., “heavy rain, two-lane highway, truck merging from the right”) are encoded via a pre-trained language encoder and injected into the diffusion UNet through cross-attention layers. The resulting attention maps A^{i,j} = softmax(Q K^T / sqrt(d))[i,j] encode the influence of text token i on spatial position j in the generated scene.

High attention weight A^{i,j} provides direct evidence that the generated scene at position j was shaped by concept i in the text description. This spatial grounding serves two policy functions. For scenario validation, engineers can verify that a generated “icy highway” scenario exhibits high attention weight on road surface regions rather than skyboxes. For regulatory documentation, the attention maps provide machine-readable evidence that scenario generation respected the input specification, satisfying the EU AI Act’s transparency obligation for AI-generated content.

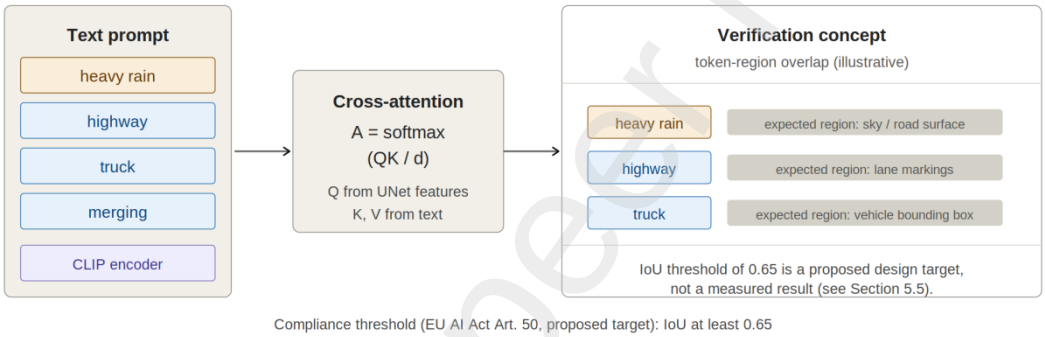

Fig. 6. Cross-attention spatial grounding mechanism. Left: text prompt tokenised into semantic tokens. Centre: cross-attention computation A = softmax(QK^T / sqrt(d)) in UNet decoder layers. Right: the proposed verification concept—checking whether each token’s attention mass falls on its expected scene region—illustrated qualitatively. No token-region IoU has been measured; the 0.65 threshold shown is a proposed design target from Section 5.5, not a reported result.

4.5 Compliance Mapping

Table 2 maps each CSD-WM output to the specific regulatory evidence artefact it satisfies. This mapping is the primary policy contribution of the paper: it translates a technical framework into actionable compliance guidance for manufacturers, certifiers, and transport policymakers.

CSD-WM output	Regulatory artefact	Instrument	Verification procedure
Factorised z_static / z_dynamic / z_env	Safety case factor independence evidence	UN ADS GTR	DCI score ≥ 0.7 on held-out factor set
DDIM inversion codes per subspace	DSSAD event reconstruction record	EU 2022/1426	Exact inversion residual < 0.01 MSE
Temporal TCAV score surface S_c^{l,τ}	Causal attribution audit trail	EU AI Act Art. 9, 11	Deletion/insertion faithfulness AUC > 0.8

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

CSD-WM output	Regulatory artefact	Instrument	Verification procedure
Cross-attention maps A^{i,j}	Scenario specification compliance evidence	EU AI Act Art. 50	Token-region IoU ≥ 0.65 on labelled scenes
Counterfactual scene generation via subspace swap	Human oversight ‘what-if’ tool	EU AI Act Art. 14, UK AV Act	Counterfactual plausibility score by domain experts

Table 2. CSD-WM interpretability outputs mapped to specific regulatory evidence artefacts.

5. Proposed Evaluation Protocol

This section specifies how CSD-WM would be evaluated empirically. No implementation of CSD-WM has been built and no experiments have been run; the framework presented in Section 4 is an architectural and conceptual proposal. We set out the protocol in detail—datasets, baselines, metrics, and target thresholds—so that the proposal is falsifiable and reproducible once implemented, and so that the compliance mapping in Table 2 can eventually be validated against real measurements rather than assumed.

5.1 Proposed Datasets and Baselines

A future evaluation should use two publicly available benchmarks. The nuScenes dataset provides urban driving sequences with multi-camera, LiDAR, and HD map annotations, suitable for assessing both generation quality and disentanglement under real-world conditions. The CARLA simulator provides controlled scenario generation with ground-truth factor labels (weather type, traffic density, road class), which is necessary for computing the DCI metrics formally, since these require oracle factor annotations that real driving datasets do not provide.

Appropriate baselines for comparison would include: (i) a standard latent diffusion world model without factorisation constraints, trained with matched architecture and compute budget, to isolate the effect of the proposed factorisation; (ii) a beta-VAE world model using the standard ELBO-based disentanglement approach, to compare against an established but architecturally simpler disentanglement method; and (iii) a representative published world model such as DriveWorld (Min et al., 2024), to situate CSD-WM's generation quality relative to the current state of the art, using that work's own reported figures rather than re-running its training.

5.2 Proposed Generation-Quality Metrics

Generation quality should be assessed using Frechet Inception Distance (FID), Precision, and Recall on held-out validation sequences, following standard practice in the generative modelling literature (Heusel et al., 2017; Kynkaanniemi et al., 2019). The central empirical question is whether the factorisation constraint imposed by CSD-WM degrades generation fidelity relative to an unconstrained baseline, and if so, by how much. A meaningful result would be a small FID cost in exchange for a substantial gain in measured disentanglement; a result in which factorisation severely degrades FID would indicate the architecture needs revision before it could be considered practical.

5.3 Proposed Disentanglement Metrics

Disentanglement should be assessed using the DCI framework (Eastwood and Williams, 2018) on CARLA, where ground-truth factor labels are available, computing Disentanglement (D), Completeness (C), and Informativeness (I) for the z_static, z_dynamic, and z_env

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

subspaces against labelled factors such as road type, weather class, and agent speed. The Mutual Information Gap (Chen et al., 2018) would provide a complementary scalar summary. We propose a target Disentanglement threshold of D ≥ 0.70 as the bar for the architecture to be considered to have achieved meaningfully separated subspaces, consistent with thresholds used elsewhere in the disentanglement literature, though this threshold itself is a design choice that should be revisited once real measurements exist.

Qualitative analysis via UMAP projection (McInnes et al., 2018) of each subspace would provide a complementary, non-quantitative check: if z_static clusters visibly by road topology and z_env clusters by weather condition without overlapping structure, this would support the quantitative DCI findings; if the projections show no such structure, this would be an early warning sign independent of the DCI computation.

5.4 Proposed Attribution Faithfulness Protocol

Attribution faithfulness for the temporal TCAV score surface should be evaluated through deletion and insertion experiments (Petsiuk et al., 2018): deletion AUC measures output degradation as the most-attributed scene regions are progressively masked, and insertion AUC measures how quickly output is recovered as those regions are revealed. Lower deletion AUC and higher insertion AUC indicate more faithful attribution. These should be compared against a gradient-saliency baseline applied to the unfactorised model to establish whether the proposed temporal TCAV approach genuinely improves on standard attribution methods rather than merely adding architectural complexity.

TCAV scores should be computed for driving concepts plausibly relevant to safety, such as icy road surface, vehicle merge conflict, pedestrian crossing, and night driving, evaluated against safety-critical prediction classes such as emergency braking. The hypothesis advanced in Section 4.3—that concepts governing scene layout will show higher attribution at high-noise (early) timesteps while concepts governing fine surface detail will show higher attribution at low-noise (late) timesteps—is a testable prediction of the architecture, not a finding; confirming or refuting it is left to future empirical work.

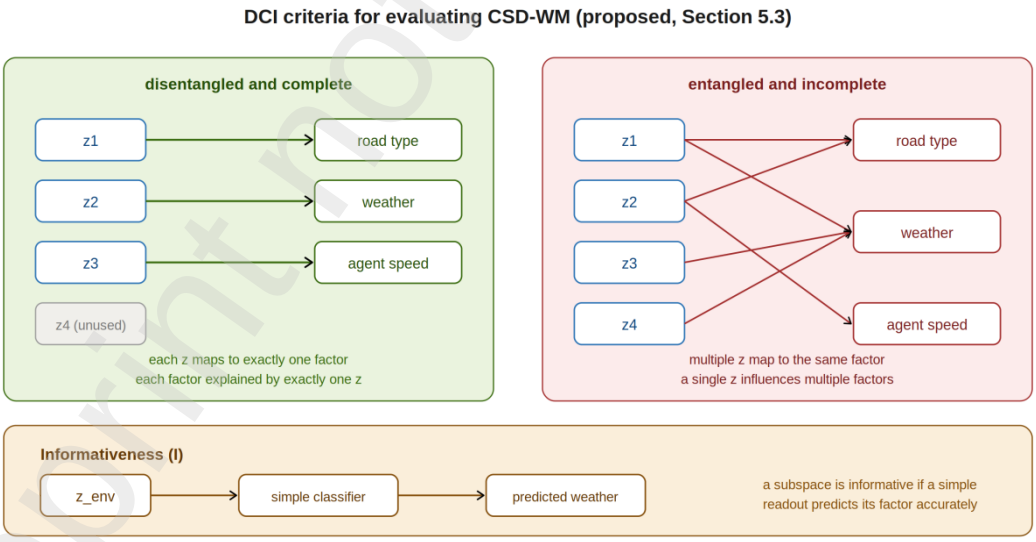

Fig. 7. DCI criteria proposed for evaluating CSD-WM (Eastwood and Williams, 2018), shown as latent-to-factor mappings. Left: a disentangled, complete mapping in which each latent dimension corresponds to exactly one scene factor. Right: an entangled, incomplete mapping in which dimensions mix multiple factors, the failure mode the factorisation objective in Section 4.1 is designed to avoid. Bottom: Informativeness, illustrated as a simple classifier reading a factor back out of its subspace. These are target definitions to be measured against a future implementation, not reported results (see Section 5.3).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

5.5 Proposed Cross-Attention Grounding Protocol

Spatial grounding quality for text-conditioned scenario generation should be assessed via token-region Intersection over Union (IoU) between cross-attention peaks and ground-truth semantic segmentation regions, across a sample of generated scenes covering a range of weather, road-type, and agent-configuration tokens. The threshold of IoU ≥ 0.65 proposed in Table 2 is a design target informed by typical attention-grounding quality reported in the image-editing literature (Hertz et al., 2023), not a measured property of CSD-WM; whether the architecture actually achieves grounding at this quality on driving scenes specifically is an open empirical question.

6. Policy Implications

6.1 From Capability Testing to Interpretability-as-Evidence

The dominant model of ADS safety assurance has been capability testing: accumulating mileage, running simulation scenarios, and demonstrating failure-rate statistics below prescribed thresholds. The UN GTR's outcome-focused, evidence-based approach represents a conceptual advance but does not by itself prescribe what constitutes evidence. The analysis in Section 3 establishes that the safety case requirement is semantically equivalent to an interpretability requirement: an evidence-based argument for ADS safety requires the ability to attribute specific model behaviours to identifiable causal factors, verify that the model responds correctly to changes in those factors, and audit the decision process retrospectively in the event of an incident.

If CSD-WM were implemented and found to satisfy multiple regulatory evidence requirements (Table 2) while maintaining generation quality competitive with non-interpretable models, this would challenge the common assumption among ADS developers that there is a fundamental performance-interpretability trade-off. Whether this holds is precisely the empirical question posed by the evaluation protocol in Section 5: a small FID cost relative to an unfactorised baseline, in exchange for a substantial gain in measured disentanglement, would support the architecture's premise; a severe FID penalty would not.

6.2 Implications for Vehicle Type Approval

EU type approval under Implementing Regulation 2022/1426 and the broader ADS Regulation requires manufacturers to submit technical documentation demonstrating DSSAD compliance and safe operational design domain (ODD) bounds. CSD-WM's factorised latent space enables a novel approach to ODD characterisation: the boundary of the z_env subspace can be defined operationally as the convex hull of environmental conditions in which the TCAV attribution structure remains stable, i.e., the same scene concepts exert the same directional influence on the same prediction classes. This "interpretability-grounded ODD" is more principled than current statistical approaches to ODD definition and directly supports the DSSAD's event occurrence analysis function.

For manufacturers seeking type approval in multiple jurisdictions, CSD-WM's compliance mapping (Table 2) provides a single technical artefact set that satisfies requirements across EU, UK, and UN GTR frameworks. Given the GTR's status as a global technical reference standard applicable in US and Canadian self-certification regimes, the framework has potential for application in non-type-approval markets without modification.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

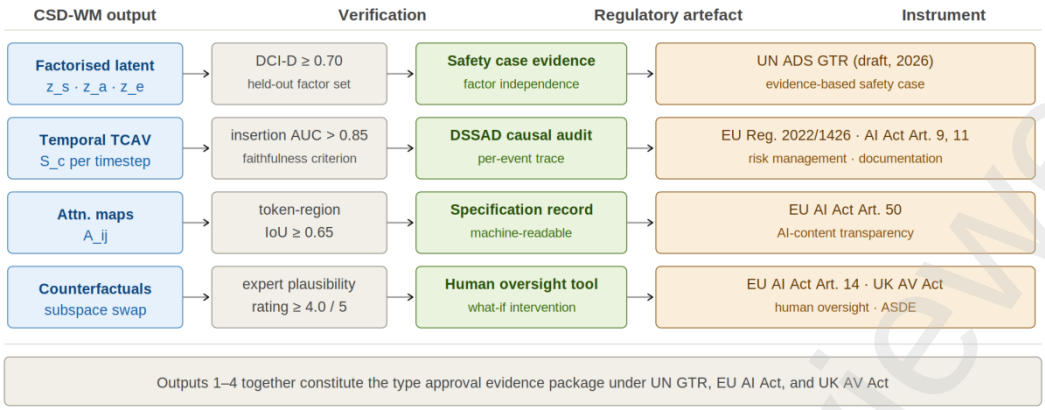

Fig. 8. Policy compliance workflow. Each CSD-WM interpretability output (left column) passes through a quantitative verification procedure (second column), produces a specific regulatory evidence artefact (third column), satisfies a named article in one or more instruments (fourth column), and serves an identified stakeholder group (right column). The bottom bar shows that the four outputs together constitute the complete evidence package required for multi-jurisdictional type approval under the draft UN GTR, the EU AI Act, and the UK AV Act.

6.3 Implications for Transport Policy and Planning

Beyond vehicle certification, interpretable world models have implications for transport planning more broadly. Transport authorities deploying ADS services in defined operational areas (e.g., urban autonomous shuttle corridors, airport automated baggage vehicles) require assurance that the simulation environment used for service design and validation accurately represents the target ODD. The cross-attention grounding provided by CSD-WM enables planners to verify that scenario generators respond correctly to ODD-defining conditions—specific road types, speed limits, intersection configurations—before committing to service design parameters derived from simulation.

The framework also enables a new class of transport policy instrument: the interpretable stress test. Rather than requiring ADS providers to demonstrate performance against a fixed scenario catalogue, regulators could specify scenarios in natural language, generate them through CSD-WM, verify semantic fidelity through cross-attention evidence, and require manufacturers to demonstrate adequate TCAV attribution stability across the generated scenario space. This approach scales indefinitely with scenario complexity and does not require physical test infrastructure.

6.4 Limitations and Caveats

Several limitations qualify the discussion above. Most fundamentally, CSD-WM has not been implemented or empirically evaluated; every quantitative claim in this paper is a proposed target or an illustrative example, not a measured result, and the architecture could fail to achieve its design goals in practice—for instance, the factorisation objective might not converge to a clean separation of subspaces, or might do so only at a generation-quality cost too severe to be practical. Second, the DCI framework requires oracle factor labels for formal computation; real driving environments do not provide such labels, so even a successful CARLA evaluation would need separate validation on real-world data before its conclusions could be trusted to transfer. Third, the compliance mapping in Table 2 offers proposed verification procedures but does not constitute legal certification; regulatory authorities retain discretion in determining what technical evidence satisfies their instruments' requirements, and this paper does not carry legal authority. Fourth, the architectural changes proposed in Section 4—independent score networks per subspace, an added total-correlation penalty—would very likely increase computational cost relative to an unfactorised baseline, though by how much can only be established once an implementation exists.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

7. Open Challenges and Future Directions

Several challenges must be addressed before CSD-WM or equivalent frameworks can be deployed in production ADS development pipelines.

7.1 Scalability of Disentanglement Evaluation

DCI metrics require oracle factor labels. Real driving scenarios involve continuously valued and partially correlated factors—weather, illumination, and road surface state are not independent—for which no ground-truth labelling infrastructure currently exists at scale. Future work should develop weak supervision approaches, potentially leveraging vision-language models for automated annotation of driving scenario factor states, to enable DCI-style evaluation on real-world driving datasets such as the Waymo Open Dataset or the Argoverse 2 dataset.

7.2 Certification-Grade Faithfulness Standards

The deletion and insertion AUC metrics proposed in Section 5 are borrowed from image classification interpretability evaluation and do not have a defined certification-grade threshold in any current ADS regulatory instrument. Collaborative standard-setting between AI safety researchers, automotive standards bodies (ISO TC22/SC32, IEEE P2846), and transport regulators is needed to establish consensus thresholds analogous to the failure-probability targets used in aviation certification.

7.3 Multi-Agent and V2X Extensions

CSD-WM decomposes the scene into static infrastructure, dynamic agent, and environmental factors, but treats agents as a collective in the z_dynamic subspace. Real safety analysis requires per-agent attribution: which specific vehicle's behaviour caused the predicted hazard? Extending the factorisation to per-agent subspaces while maintaining generation quality is a non-trivial modelling challenge, particularly for dense urban scenarios with many interacting agents. Connected and automated driving contexts (V2X) introduce additional shared-state information that should ideally have its own latent subspace with corresponding attribution semantics.

7.4 Standardised Benchmarks for World Model Interpretability

The lack of standardised benchmarks for generative model interpretability, identified by Chakravarthy and Roy as an open challenge for GAN and diffusion interpretability generally, applies with equal force to world models. The research community would benefit from a dedicated benchmark combining generation quality, disentanglement, attribution faithfulness, and regulatory compliance evidence metrics, evaluated on a shared dataset with ground-truth factor labels. Development of such a benchmark in collaboration with automotive OEMs and transport authorities represents a high-priority infrastructure investment for the field.

8. Conclusion

This paper has presented the case that interpretability in autonomous driving world models is a regulatory necessity, not merely a technical preference. The draft UN Global Technical Regulation on Automated Driving Systems, recommended by GRVA for WP.29 adoption in 2026, and the EU AI Act (enforceable August 2026) together establish a regulatory direction in which the opaque latent representations of current state-of-the-art world models are structurally non-compliant. The proposed CSD-WM framework addresses this gap through architecturally enforced causal scene decomposition, temporal TCAV attribution, cross-attention spatial grounding, and exact DDIM inversion—each mapped to specific regulatory evidence requirements.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

This paper has presented an architecture and a proposed evaluation protocol; it has not presented experimental results. The central claim is conceptual: that architecturally enforced disentanglement, rather than post-hoc explanation alone, offers a promising route to interpretability outputs that map onto concrete regulatory evidence requirements (Table 2). Whether CSD-WM achieves competitive generation quality alongside meaningful disentanglement is an open empirical question that the protocol in Section 5 is designed to answer once an implementation exists.

As generative world models become embedded in both ADS development pipelines and transport planning practice, the demand for interpretable, auditable, and accountability-ready AI will intensify. The framework and regulatory analysis presented here offer a starting point for meeting that demand, but they require empirical validation before any compliance, performance, or policy claim attached to CSD-WM should be relied upon. The open challenges identified in Section 7, together with the evaluation protocol in Section 5, define the next steps.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. In Proceedings of NeurIPS, vol. 31.

[2] Da, L., Chen, T., Li, Z., Bachiraju, S., Yao, H., Li, L., Dong, Y., Hu, X., Tu, Z., Wang, D., Zhao, Y., Zhou, B., Pendyala, R., Stabler, B., Yang, Y., Zhou, X., and Wei, H. (2025). Generative AI in Transportation Planning: A Survey. arXiv:2503.07158.

[3] Eastwood, C. and Williams, C. K. I. (2018). A framework for the quantitative evaluation of disentangled representations. In Proceedings of ICLR Workshop.

[4] European Commission. (2022). Commission Implementing Regulation (EU) 2022/1426 laying down rules for the type-approval of the automated driving system (ADS) of fully automated vehicles. Official Journal of the European Union, L221.

[5] European Parliament and Council. (2024). Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). Official Journal of the European Union.

[6] Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., and Li, H. (2024). Vista: A generalizable driving world model with high fidelity and versatile controllability. In Advances in NeurIPS, vol. 37, pp. 91560–91596.

[7] Ha, D. and Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In Advances in NeurIPS, vol. 31.

[8] Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. (2019). Learning latent dynamics for planning from pixels. In Proceedings of ICML, pp. 2555–2565.

[9] Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. (2025). Mastering diverse control tasks through world models. Nature, 640(8059), 647–653.

[10] Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. (2020). GANSpace: Discovering interpretable GAN controls. In Advances in NeurIPS, vol. 33, pp. 9841–9850.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792

[11] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. (2023). Prompt-to-Prompt image editing with cross attention control. In Proceedings of ICLR.

[12] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in NeurIPS, vol. 33, pp. 6840–6851.

[13] Hu, A., et al. (2023). GAIA-1: A generative world model for autonomous driving. arXiv:2309.17080.

[14] UNECE Working Party on Automated/Autonomous and Connected Vehicles (GRVA). (2026). Draft UN Global Technical Regulation on Automated Driving Systems (ADS), adopted by GRVA in January 2026 and recommended for adoption by WP.29. United Nations Economic Commission for Europe, Geneva.

[15] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viégas, F., and Sayres, R. (2018). Interpretability beyond classification accuracy: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of ICML.

[16] Kim, J. and Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In Proceedings of ICCV, pp. 2942–2950.

[17] Li, Y., Fan, L., He, J., Wang, Y., Chen, Y., Zhang, Z., and Tan, T. (2024). Enhancing end-to-end autonomous driving with latent world model. arXiv:2406.08481.

[18] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in NeurIPS, vol. 30.

[19] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.

[20] Min, C., Zhao, D., Xiao, L., Zhao, J., Xu, X., Zhu, Z., Jin, L., Li, J., Guo, Y., Xing, J., and Wang, J. (2024). DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving. arXiv:2405.04390.

[21] Petsiuk, V., Das, A., and Saenko, K. (2018). RISE: Randomised input sampling for explanation of black-box models. In Proceedings of BMVC.

[22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of CVPR, pp. 10684–10695.

[23] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2), 336–359.

[24] Shen, Y., Gu, J., Tang, X., and Zhou, B. (2020). Interpreting the latent space of GANs for semantic face editing. In Proceedings of CVPR, pp. 9243–9252.

[25] Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In Proceedings of ICLR.

[26] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of ICML, pp. 3319–3328.

[27] UK Government. (2024). Automated Vehicles Act 2024. Her Majesty's Stationery Office, London.

[28] Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., and Lu, J. (2023). DriveDreamer: Towards real-world-driven world models for autonomous driving. arXiv:2309.09777.

[29] Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., and Lu, J. (2023). OccWorld: Learning a 3D occupancy world model for autonomous driving. arXiv:2311.16038.

[30] Zheng, Y. et al. (2025). World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. arXiv:2507.00603.

[31] Zhang, L., Xiong, Y., Yang, Z., Casas, S., Hu, R., and Urtasun, R. (2023). Copilot4D: Learning unsupervised world models for autonomous driving via discrete diffusion. arXiv:2311.01017. Published at ICLR 2024.

[32] Jia, F. et al. (2026). DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving. arXiv:2602.06521.

[33] Wayve. (2026). A global regulatory breakthrough for assisted and automated driving. Wayve Technical Blog, February 2026. https://wayve.ai/thinking/a-global-regulatory-breakthrough-for-assisted-and-automated-driving/

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7032792